Walk in Others' Shoes with a Single Glance: Human-Centric Visual Grounding with Top-View Perspective Transformation

Yuqi Bu[1,2,3]*, Xin Wu[2,3], Zirui Zhao[4], Yi Cai[2,3]†, David Hsu[4], Qiong Liu[2]
[1]School of Artificial Intelligence, Shenzhen Polytechnic University
[2]School of Software Engineering, South China University of Technology
[3]Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
[4]School of Computing, National University of Singapore

Abstract

Visual perspective-taking, an ability to envision others' perspectives from a single self-perspective, is vital in human-robot interactions. Thus, we introduce a human-centric visual grounding task and a dataset to evaluate this ability. Recent advances in vision-language models (VLMs) have shown potential for inferring others' perspectives, yet are insensitive to information differences induced by slight perspective changes. To address this problem, we propose a top-view enhanced perspective transformation (TEP) method, which decomposes the transition from robot to human perspectives through an abstract top-view representation. It unifies perspectives and facilitates the capture of information differences from diverse perspectives. Experimental results show that TEP improves performance by up to 18%, exhibits perspective-taking abilities across various perspectives, and generalizes effectively to robotic and dynamic scenarios. ‡

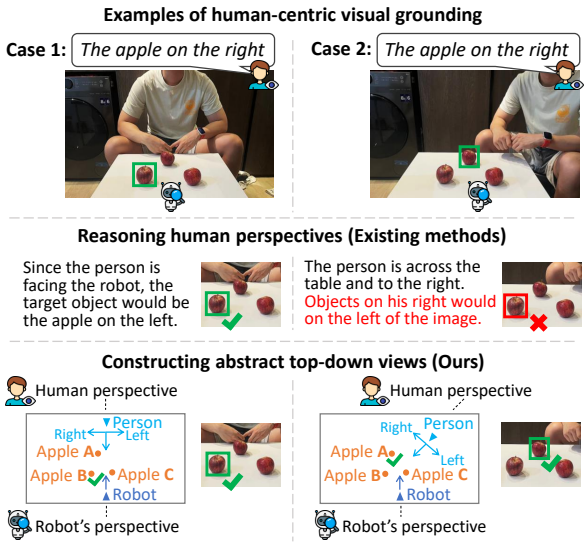


Figure 1: Examples of human-centric visual grounding. Existing methods identify the same object regardless of perspective shifts, revealing an insensitivity problem. Our proposed method constructs a unified top view to capture information differences across perspectives.

1 Introduction

Human-robot interaction (HRI) aims to enable natural and efficient collaboration between humans and robots. A critical aspect of HRI is interpreting human intentions, particularly reconciling human egocentric instructions with a robot's allocentric operations (Li et al., 2016; Trafton et al., 2005; Chen et al., 2023; Berlin et al., 2006). Failure to align these perspectives may result in misinterpretation and dangerous errors. For example, when a human engaged in a task says, "Hand me the tool on my left," a robot that ignores the speaker's perspective might retrieve the wrong, potentially hazardous tool. Thus, aligning human and robotic perspectives is essential for effective HRI.

Research in cognitive psychology (McGee, 1979; Langdon and Coltheart, 2001) reveals that humans possess a visual perspective-taking ability, allowing them to infer others' perspectives (*i.e.,* walk in others' shoes) solely from an egocentric perspective (*i.e.,* with a single glance). This finding raises a critical yet underexplored question: How can robots acquire this ability based on their single perspective? Motivated by this challenge, we propose a human-centric visual grounding (HVG) task that requires robots to interpret human-centric instructions and identify corresponding objects from robot perspectives. The difficulty stems from the robot's limited perspective, compelling it to simulate human perspectives using first-person visual data, which is valuable in resource-constrained environments (Neuman et al., 2022; Ma et al., 2016). HVG differs from conventional visual grounding (Yu et al., 2016; Bu et al., 2023a) (also known as re-

*This work was partially done while Yuqi Bu was a visiting student at the National University of Singapore.
†Corresponding author: Yi Cai (ycai@scut.edu.cn).
‡The code is available at https://github.com/Buki2/TEP.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26904–26923

ferring expression comprehension), which assumes the same perspective for speakers and operators.

Recent advances in large language models (LLMs) (Achiam et al., 2023) and vision-language models (VLMs) (Liu et al., 2023) showcase perspective-taking abilities, yet they remain **insensitive to information differences induced by slight perspective changes**. As shown in Fig. 1, Case 1, a VLM, GPT-4V, correctly infers spatial reversals when human and robot perspectives are directly opposite. However, in Case 2, a slight shift in the human perspective alters the target object, yet this model still assumes a simple spatial reversal, leading to incorrect grounding. This highlights the insensitivity of existing models to nuanced spatial reference changes across perspectives.

In human cognition, visual perspective-taking involves an allocentric reference frame, which is a mental abstraction based on objective environmental landmarks (Klatzky, 1998; Langdon and Coltheart, 2001; Tversky and Hard, 2009). Inspired by this mechanism, we propose generating an abstract top-down view of a visual scene as an allocentric reference frame to address the problem of insensitivity to perspective-induced information differences. This view unifies human and robot perspectives into a shared space and abstracts visual scenes by simplifying object locations and human orientations, which enables observations of spatial information from different perspectives. As shown in Fig. 1, our top-down view clarifies spatial references of human perspectives, allowing the robot to correctly identify Apple B and Apple A as the target objects in Cases 1 and 2, respectively.

We propose a top-view enhanced perspective transformation (TEP) method, which bridges the transition from robot to human perspectives by an intermediate top-view representation. Recognizing that HVG alters the conventional vision-language alignment (*e.g.,* equating the image's left with the text "left"), supervised learning on HVG data can induce confusion. To circumvent task-specific training, a grounding decomposer elicits LLMs to interpret instructions and decompose tasks into visual modules (*e.g.,* object retrieval and spatial reasoning). Subsequently, to address the insensitivity problem, a spatial reasoning module utilizes VLMs-based symbolic reasoning to construct an abstract top-view representation. This representation includes generated hints that correspond to the intended spatial information, thereby enhancing cross-perspective reasoning. Finally, to tackle the scarcity of multi-perspective data, we introduce an InterRef dataset, comprising 1K test samples. Experiments show that TEP improves performance by up to 18% across seven types of perspectives and generalizes to robotic and dynamic scenarios.

Our contributions are summarized as follows:

- We propose a task and dataset for perspective-taking in visual grounding, which requires inferring object information from others' perspectives based on a single self-perspective.
- To address the insensitivity to perspective-induced information differences, we propose the TEP method, which enhances cross-perspective reasoning by translating spatial information through a top-view representation.
- Experiments show that TEP improves performance by up to 18%, exhibits perspective-taking abilities across perspectives, and generalizes in robotic and dynamic scenarios.

2 Related Work

2.1 Perspective-Related Tasks

Visual perspective-taking is crucial for effective HRI (Li et al., 2016; Trafton et al., 2005). Previous robotic tasks mainly focus on perception learning with sensory cues. These tasks determine human perspectives by scene descriptions (Huang et al., 2022a), multi-view images of the entire scene (Dogan et al., 2020; Huang et al., 2022a), or robot movement (Pramanick et al., 2022). Grounding is fundamental for HRI (Reich and Schultz, 2024; Xiao et al., 2024). While most grounding tasks (Yu et al., 2016; Bu et al., 2023a,b) are self-centric, some involves perspectives. For instance, 3D grounding (Achlioptas et al., 2020; Guo et al., 2023) involves text input that specifies perspectives, and embodied grounding (Shi and Yang, 2022; Islam et al., 2023) uses gestures to indicate perspectives. However, these tasks rely on multiple viewpoints (*e.g.,* point cloud) or additional context (*e.g.,* scene description). Heavy reliance on sensory cues may hinder scalability in open-world environments, such as limited movement space. These task settings diverge from the human-like ability to envision others' perspectives using a self-perspective.

This paper explores reasoning about human perspectives using only a robot's perspective. This resource-limited setting prioritizes reasoning ability and enhances generalization to rare scenarios, minimal data needs, and better interpretability.

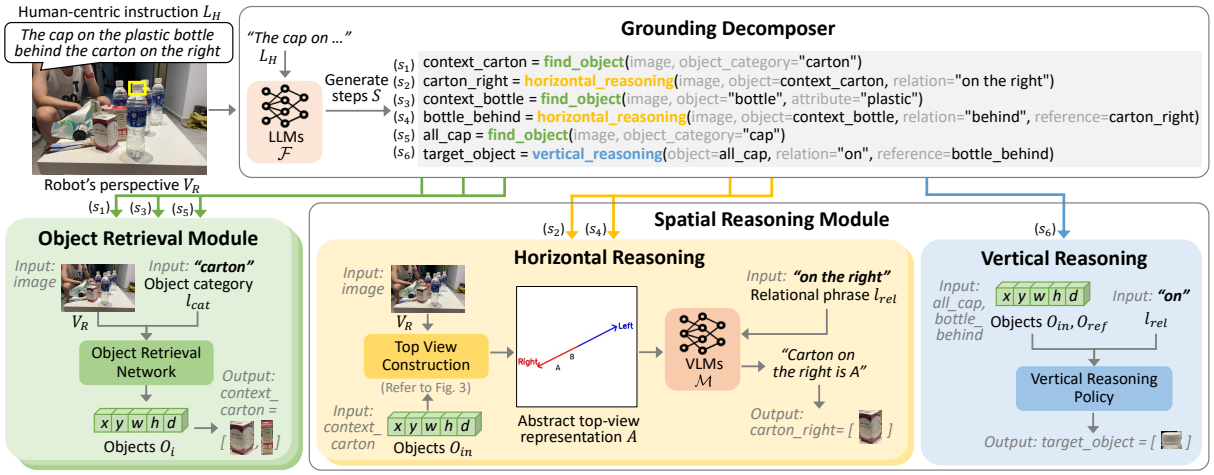


Figure 2: The framework of TEP. A grounding decomposer breaks down the task into steps. These steps engage an object retrieval module to extract visual information and a spatial reasoning module to discern spatial relationships.

Moreover, the ego-exo alignment task (Grauman et al., 2024; Huang et al., 2024, 2025) has potential applications in HRI, such as predicting human perspectives from robot views. However, it mainly targets human pose prediction in scenes with few or dissimilar objects. In contrast, visual grounding involves dense, similar objects, making prediction under such conditions more challenging.

2.2 Perspective-Taking Ability of Models

Existing perspective-taking methods (Dogan et al., 2020; Huang et al., 2022a; Pramanick et al., 2022; Shi and Yang, 2022; Islam et al., 2023) require information from ample sensors, rendering them unfit for this resource-limited task. Besides, conventional grounding methods (Kamath et al., 2021; Yan et al., 2023; Wang et al., 2022) center on robot's perspectives and require extensive task-specific data to learn grounding from human perspectives, which is laborious and time-consuming. Moreover, existing exo-to-ego methods (Liu et al., 2024; Xu et al., 2025) require multiple exo views for full scene coverage and focus on visual synthesis and rendering.

Foundation models (*e.g.,* LLMs) (Achiam et al., 2023; Liu et al., 2023; Li et al., 2023) are notable for zero-shot reasoning abilities and broad knowledge. Recent trends leverage them to decompose long-horizon tasks (Huang et al., 2022b; Wu et al., 2024), analyze multimodal information (Kamath et al., 2023; Li et al., 2024; Fan et al., 2024), and reason about physical environments (Sclar et al., 2023; Du et al., 2024). For perspective-taking tasks, these models infer human perspectives via various techniques such as CoT (Wei et al., 2022), while they suffer from insensitivity to information differences across perspectives. This may stem from perspective imbalances in data, *i.e.,* there is an overabundance of robot-centric instructions and face-to-face HRI, thus marginalizing certain scenes such as a diagonal perspective. Thus, we present a top-view representation to unify perspectives, capture information differences, and enhance the perspective-taking abilities of foundation models.

3 Method

3.1 Problem Statement

Assume a scene with a robot, a human, and a set of objects O arranged on a table, with both the human and robot surrounding and facing the table. The human and robot have observations from their perspectives: V_H and V_R. When the human instructs the robot to pick up a target object $O_{tgt} \in O$, they provide instruction (or referring expression) L_H based on V_H. Then, the robot should ground O_{tgt} in its view V_R. Thus, the goal of HVG is defined as $P(O_{tgt}|V_R, L_H)$. The instructions manifest as text detailing object categories, attributes, and relationships, while the robot observations are 2D images $V_R \in \mathbb{R}^{H\times W\times 3}$. It is assumed that all mentioned objects are visible and specified by name.

3.2 Overall Architecture

The framework of TEP is shown in Fig. 2. Its core is an abstract top-view representation for cross-perspective spatial reasoning. It is based on a hypothesis that by establishing the spatial relationship mappings from V_R to a top view V_T and from V_T to V_H, we can derive the V_R-V_H mapping.

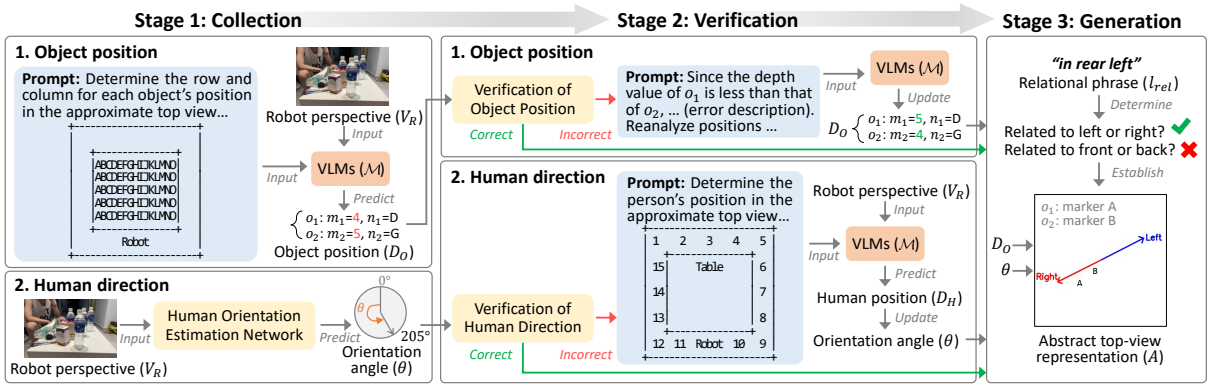


Figure 3: The construction process of the abstract top-view representation in TEP.

3.3 Grounding Decomposer

Understanding human intention within instructions is a prerequisite for the HVG task. To achieve this, the grounding decomposer F_{GND} leverages the language comprehension ability of LLMs to parse an instruction and decompose it into a series of grounding steps $S = \langle s_1, s_2, ..., s_{|S|} \rangle$. These steps are predicted through $P(S|L_H; \mathcal{F}, \rho)$, where $\mathcal{F}$ denotes an LLM, GPT-4, and ρ serves as a prompt.

The grounding steps serve to progressively identify target objects through an object retrieval module F_{RETR} and a spatial reasoning module with horizontal reasoning F_{HORIZ} and vertical reasoning F_{VERT} functions. This module generates code for each step and then executes the code to derive the results. In particular, each step s_i is represented as a line of code and produces a set of objects $O_i = \{o_j\}_{j=1}^{|O_i|}$ that satisfy the constraints specified within this step. The module employs steps S as a bridge, where the LLM generates S, and the visual models execute it based on V_R. Upon completion, the target object O_{tgt} is obtained, represented as $P(O_{tgt}|V_R, S)$.

Compared to free-form text, the logical structure of code allows flexibility in representing steps. Moreover, we present a context-first strategy that prioritizes identifying context objects and associated reasoning before identifying the target object, thus reducing computations in task decomposition.

3.4 Object Retrieval Module

To extract visual information, this module retrieves objects O_i in V_R that match the object category l_{cat} and attributes l_{attr} constrained in a given step s_i and then extracts positions $\{x, y, w, h\}$ and depth values d of retrieved objects. This process is denoted as $O_i = F_{RETR}(V_R, l_{cat}, l_{attr})$, where each object o_j in O_i contains visual information of $o_j = \{x, y, w, h, d\}$. This module utilizes a text-based object retrieval network, GLIP-Large (Li et al., 2022), to extract object bounding boxes in V_R and a monocular depth estimation network, MiDaS-DPT-Large (Ranftl et al., 2022), to predict depth values within object regions.

3.5 Spatial Reasoning Module

In this module, each step s_i contains a relational constraint $K = (O_{in}, l_{rel}, O_{ref})$. Here, O_{in} and O_{ref} denote sets of candidate target objects and context objects, respectively, while l_{rel} is a relational phrase describing their spatial relationships. Given a constraint K, this module identifies objects $O_i \subseteq O_{in}$ that satisfy l_{rel} relative to O_{ref}. In cases where the constraint excludes context objects (*e.g.,* "book on the left"), $O_{ref} = \varnothing$.

We split spatial reasoning into horizontal (*e.g.,* left/right) and vertical (*e.g.,* above/below) functions. In the task considered in this paper, since humans and robots typically stand upright in a fixed top-to-bottom orientation, horizontal relationships among objects vary with the observer's positions, while vertical relationships are less affected. Thus, vertical reasoning generally does not require human orientations in the top view, allowing us to simplify it by eliminating the top view mechanism.

Horizontal Reasoning. F_{HORIZ} construct an abstract top-view representation A to analyze horizontal relationships among objects across perspectives. Since object positions are more essential than appearances for determining spatial relationships, we propose an abstract representation for V_T. It contains object indexes as markers for their positions and arrows that indicate directions of human perspectives. To avoid interference from unrelated objects, we construct an A for O_{in} and O_{ref} at each

step, ensuring no overlap at identical positions.

The construction of A follows three stages: information collection, verification, and generation (Fig. 3). In the first stage, a VLM $\mathcal{M}$ infers the positions D_O of objects O_{in} and O_{ref} in V_T, denoted as $P(D_O|V_R, O_{in}, O_{ref}; \mathcal{M}, \rho)$. Inspired by the symbolic reasoning abilities of VLMs (Mirchandani et al., 2023) and the visualization of multimodal data (Wang et al., 2025), we design a symbol-based grid to approximate the spatial structure of V_T. It includes a central rectangle representing the table, divided into $M = 5$ rows and $N = 15$ columns. The VLM analyzes object relationships in V_R and assigns grid coordinates (m_j, n_j) for each object o_j. The grid serves as an abstraction to represent a surface that holds objects, chosen for its simplicity in delineating the surface boundaries. Other shapes, such as circles, introduce challenges in symbolic depiction due to the complexity of curved lines compared to straight edges. Additionally, a monocular human orientation estimation network (Wu et al., 2020) predicts body orientation angle $\theta \in [0°, 360°)$ from cropped human regions in V_R. This network uses a visual backbone and prediction head to estimate angles, which are then used to approximate human direction in V_T. We use body cues to predict orientation, as 78.70% of samples lack discernible facial features.

The second stage validates collected information using predefined rules. It assesses whether the relative positions of objects are consistent between V_T and V_R. For instance, objects with smaller depth values in V_R, *i.e.,* closer to robots, should appear lower in V_T. If inconsistent, a VLM adjusts positions by predicting $P(D_O|V_R, O_{in}, O_{ref}, l_{err}; \mathcal{M}, \rho)$, where l_{err} describes errors. Similarly, human orientation verification ensures valid human-table relationships in both views. If incorrect, a VLM identifies human positions D_H in a symbol-based grid with candidate positions around a rectangular table, *i.e.,* $P(D_H|V_R, o_{hmn}, o_{tbl}; \mathcal{M}, \rho)$, where o_{hmn} and o_{tbl} represent object information of human and table. Then, the direction from D_H to the center of V_T is used to update the initial orientation angle θ.

The third stage generates a top view using verified information. First, markers for objects O_{in} and O_{ref} are placed at D_O. Then, a spatial reference system is established for human perspectives in V_T by analyzing the relational phrase l_{rel} to determine directional arrows (*e.g.,* left/right, front/back). Finally, the generated top view A includes object markers and directional arrows, as shown in Fig. 3.

Once A is generated, F_{HORIZ} identifies objects $O_i \subseteq O_{in}$ that conform to a given constraint K at step s_i. This is achieved by eliciting a VLM $\mathcal{M}$ to predict $P(O_i|A, O_{in}, l_{rel}, O_{ref}; \mathcal{M}, \rho)$. The VLM used in this function is GPT-4V. Evaluations of various VLMs are presented in Section 4.3.

Vertical Reasoning. This function F_{VERT} identifies objects $O_i \subseteq O_{in}$ that satisfy a relational constraint K at step s_i, formulated as $O_i = F_{VERT}(O_{in}, l_{rel}, O_{ref})$. We construct a program based on a vertical reasoning policy. It uses y-coordinates and depth values of objects to determine spatial relationships, following a principle that larger y-coordinates and depth values correspond to objects positioned lower in V_H.

4 Experiment

4.1 Experimental Setup

In TEP, we employ GPT-4 as the LLM $\mathcal{F}$ and GPT-4V as the VLM $\mathcal{M}$. For implementation details and prompts, please refer to Appendices A and F.

Dataset. Due to the lack of perspective diversity and available data in existing datasets (Li et al., 2016; Dogan et al., 2020), we construct a new dataset, InterRef, for the HVG task. It contains 1,069 human-centric instructions corresponding to 130 real-world images, providing a range of perspectives. The dataset is split into a validation set of 50 samples and a test set of 1,019 samples.

Data Collection: We capture 84 robot scene images and collect 46 images from the MSCOCO dataset (Lin et al., 2014). In each image, a person is around a table with multiple objects. For object annotations, we use object detection results for robot scene images and original annotations for the COCO images. We manually annotate each expression to uniquely refer to a target object. These expressions are recorded on-site for robot scene images and imagined from the perspective of the person in the COCO images.

Perspective Classification: We classify samples based on the differences in perspectives between humans and robots. Before this, we label fifteen human positions around a table, expanding the typical setup (Dogan et al., 2020) of eight perspectives (*i.e.,* four orthogonal and four diagonal) to include seven additional intermediate perspectives. This expansion allows for capturing a wider range of scenarios. Then, we classify perspectives into four directional categories: opposite (opp.), left, right,

Method	Test set	Directional perspectives				Angular perspectives		
		Opp.	Left	Right	Same	Orthog.	Diag.	Other
Supervised visual grounding models								
MDETR (ResNet-101)	10.79	8.65	11.64	13.14	11.64	10.21	12.03	10.53
UNINEXT (huge)	12.37	12.20	13.82	13.66	3.42	13.03	10.37	12.96
OFA (large)	20.51	12.64	23.27	24.74	26.03	17.25	19.09	23.08
Vision-language models								
Gemini (1.0-pro-vision)	21.10	14.61	22.03	19.12	28.13	18.52	15.91	22.64
LLaVA (v1.6-34B)	23.53	19.10	22.03	25.00	34.38	25.93	22.73	22.64
GPT-4V	25.81	20.40	30.18	23.71	30.82	23.24	20.33	29.96
GPT-4V w/ orientation	31.50	32.15	32.36	26.80	32.19	31.69	26.56	33.81
Manual perspective transformation method								
Rephrase + OFA	40.53	49.00	32.73	39.69	25.34	47.54	36.93	38.26
Rephrase + GPT-4V	47.89	49.00	48.36	42.01	41.10	57.04	36.93	47.98
TEP (Ours)	**66.73**	**65.63**	**69.09**	**68.30**	**58.90**	**66.20**	**65.15**	**67.81**

Table 1: Comparison with baseline models on the test set of the InterRef dataset.

and same, as well as three angular perspectives based on direct angle: orthogonal (orthog., *e.g.,* immediate left), diagonal (diag., *e.g.,* front-left), and other. For more data details, refer to Appendix B.

Evaluation Metrics. Perspective-taking ability is evaluated by grounding task accuracy (Yu et al., 2016), defined as the average number of correct predictions (*i.e.,* IoU greater than 0.50).

TEP enables perspective-taking analysis by evaluating object positions (Acc@OP) and human directions (Acc@HD) in generated top views. Object positions are correct if all object pairs in a top view exhibit accurate and distinguishable front-back and left-right correlations. We manually evaluate 101 samples due to annotation complexity. Human direction accuracy is measured by the angle between the forward direction and the ground truth (from annotated human positions to top-view centers), considering it accurate if within ± 30 degrees.

Baseline Models. We evaluate state-of-the-art grounding models, MDETR (Kamath et al., 2021), UNINEXT (Yan et al., 2023), and OFA (Wang et al., 2022), as they have shown dataset generalizability. Moreover, we evaluate popular VLMs, Gemini, LLaVA (Liu et al., 2023), and GPT-4V (Achiam et al., 2023) using the same parameters as in TEP. Their CoT prompts provide bounding boxes and depth values of all objects. Gemini and LLaVA are evaluated on 20% of the test set.

We present a manual perspective transformation method that converts human-centric instructions into robot-centric ones using rules. This method divides angles from (Wu et al., 2020) into four sections: the person is on the opposite, left, right, or the same side. Spatial phrases in expressions are then rephrased to match the robot's perspective. For example, when a person is on the opposite, "left" becomes "right". The rephrased expressions are processed by OFA and GPT-4V for grounding.

4.2 Evaluation of Perspective-Taking Ability

Overall Performance. The results in Table 1 show that supervised visual grounding models struggle with the HVG task. This may be attributed to their focus on robot-centric grounding and lack of perspective-taking. Among them, OFA excels, demonstrating superior generalizability and outperforming the others on this partially unseen dataset.

In addition, VLMs exhibit certain perspective-taking abilities and achieve over 20.00% accuracy, while this potential remains underutilized. GPT-4V slightly outperforms LLaVA, whereas augmenting GPT-4V's input with human orientation information boosts its performance by 5.69%. This highlights the value of such information in improving reasoning about human perspectives, in line with our decomposition method that explicitly identifies and integrates human orientation. For few-shot experiments, refer to Appendix C.

Moreover, the manual perspective transformation methods show substantial improvements of 20.02% and 22.08% over OFA and GPT-4V, respectively. However, these rule-based methods are criticized for their inflexibility, as well as an inability to discern slight changes in perspectives.

TEP outperforms these strong baselines by

18.84%, demonstrating its perspective-taking ability to envision information in human perspectives based on a single robot's perspective. Nonetheless, compared to the nearly 90.00% accuracy in conventional grounding tasks (Wang et al., 2022), the results underscore this task's complexity. Please refer to Appendix D for more qualitative results.

Furthermore, we evaluate an end-to-end inference time on a Tesla T4 and Intel Xeon@2.00GHz server. TEP runs at 13.00s per image, with 5.27s (40.54%) dedicated to top-view construction and 3.90s (30.00%) for the object retrieval module, while GPT-4V w/ orientation takes 16.89s due to analyzing all objects in each image. Besides, in TEP, the average input and output tokens are 476.2 and 177.6 for LLMs, and 1600.4 and 525.0 for VLMs. The baseline model averages 455.4 input tokens and 519.9 output tokens for VLMs.

Sensitivity of Directional Perspectives. Table 1 shows most visual grounding models and VLMs perform worse on opposite-side human interactions but better on same-side interactions. The reason is that same-side samples align with conventional grounding scenarios, which are familiar contexts for baselines. After incorporating orientation, GPT-4V's opposite-side performance improves significantly, indicating that human orientation enhances its perspective-taking ability for such cases.

Moreover, the manual perspective transformation methods perform better on opposite-side interactions, due to mirrored spatial relationships simplifying transformation. However, integrating this method into OFA slightly reduces same-side performance, since orientation prediction errors cause instructions to be rephrased to mismatched robot perspectives, revealing an inflexibility issue.

Furthermore, TEP outperforms all baselines and demonstrates a consistent accuracy improvement of over 15.00% across all perspectives, underscoring its generalization and perspective sensitivity. Further experiments show that in same-side samples, humans often appear with their backs turned to the camera, making orientation prediction more difficult (*e.g.,* Acc@HD is 54.54% vs. over 70% in other categories). This challenge likely hampers subsequent cross-perspective spatial reasoning, reducing overall accuracy in this category.

Sensitivity of Angular Perspectives. Table 1 reveals that baseline models struggle with diagonal perspectives, highlighting the inherent challenges of these scenarios. GPT-4V with human orientation performs worst in such perspectives, exposing its insensitivity to information differences from these perspectives. Conversely, TEP achieves the most significant improvements in these scenarios, **proving its efficacy in mitigating this insensitivity problem.** Additionally, incorporating rephrasing methods into OFA and GPT-4V enhances their performance in orthogonal perspectives, owing to the transformation rules tailored to such scenarios.

Grounding decomposer	Abstract top-view representation	Accuracy
✗	✗	31.50
✓	✗	46.52 (+15.02)
✓	✓	**66.73** (+20.21)

Table 2: Ablation study on variants of the TEP method.

Block	Method	Accuracy
A. Object position	Text representation	45.34
	Box representation	57.90
	Point representation	**66.73**
B. Human direction	Position marker	43.67
	Orientation arrow	47.69
	Relevant direction	**66.73**

Table 3: Experiments on top-view representation.

4.3 Evaluation of Top-View Representation

Significance of Top View. Table 2 presents ablation studies evaluating the effectiveness of the abstract top-view representation for perspective-taking in TEP. The baseline (*i.e.,* GPT-4V w/ orientation) lacking both the grounding decomposer and the top-view representation achieves an accuracy of 31.50%. Incorporating the grounding decomposer, which is identical to the TEP setup for grounding decomposition and object retrieval, improves performance by 15.02%. In addition, integrating the top-view representation yields an additional gain of 20.21%. These results suggest that directly applying VLMs for all steps after decomposition is suboptimal and the top-view representation plays a critical role in enhancing overall performance.

Variations in Top-View Representation. In Table 3 Block A, we substitute the point representation of object positions in TEP with text or box representations, as shown in Fig. 4. The text representation, generated by GPT-4V using V_R to describe spatial information in a top view, yields the lowest accuracy due to inadequate granularity and intuitiveness for spatial reasoning. The box representation is 8.83% less accurate than the point

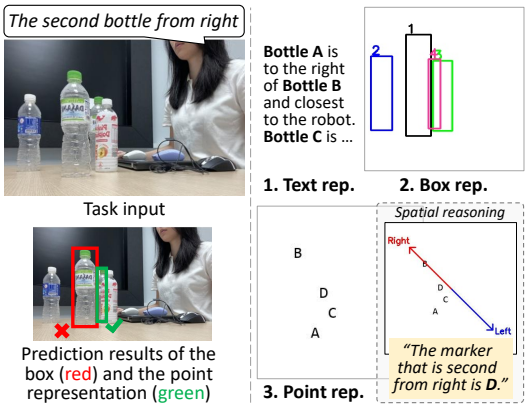


Figure 4: Visualization of top-view representation.

Method	Acc@OP	Acc@HD	HVG
Box center	63.37	67.03	57.70
Ours w/o ver.	69.31	82.34	62.90
Ours	**78.22**	**83.02**	**66.73**

Table 4: Experiments on top-view construction method.

representation, as 2D boxes omit depth, distorting front-back relationships in the top view. Fig. 4 shows the box representation arranges objects side-by-side, lacking front-back relationships, causing object 1 to be misidentified as the target. The point representation clearly depicts left-right and front-back relationships, facilitating effective reasoning.

Block B in Table 3 shows that representing human positions with markers or using arrows to denote human orientation is suboptimal. Neither method provides sufficient information for reasoning about object relationships from human perspectives, thereby requiring further determining the reference frames of human direction. In contrast, TEP's relevant directions provide arrows indicating the directions related to the intended spatial relationships, streamlining spatial reasoning.

Variations in Top-View Construction Method. In Table 4, the first row uses the center of object bounding boxes in V_R as the point representation for object positions and the center of the human bounding box to determine the human direction. This method performs worst, suggesting that box information is insufficient in depicting spatial information in V_T. In contrast, our method integrates bounding boxes with depth information for a more accurate representation.

The second row shows that without the verification stage (ver.), the top-view construction process of TEP achieves 69.31% accuracy for object positions in generated top views and 82.34% accuracy for human directions. After applying the verification stage, accuracies improve by 8.91% and 0.68%, because of eliminating local errors in the initial representation. Besides, of the errors found by the verification stage, accuracies for identifying actual errors are 66.67% for object positions and 100.00% for human directions, with correction accuracies of 70.83% and 65.38%, respectively.

Method	Accuracy
VQA model (BLIP-2)	41.12
VLMs (LLaVA)	58.42
VLMs (GPT-4V)	**66.73**

Table 5: Experiments on spatial reasoning methods.

Figure 5: Experiment setting in real robot environments.

Variations in Top-View Reasoning Method. Table 5 evaluates three models for spatial reasoning on the abstract top-view representation. First, we formalize the spatial reasoning as a visual question answering (VQA) task and pose questions to a powerful VQA model, BLIP-2 (Li et al., 2023), such as "which marker is on the left". It achieves an accuracy of 41.12%, indicating limited effectiveness in abstract diagram reasoning. In contrast, LLaVA and GPT-4V bring about a 17.30% improvement due to enhanced reasoning abilities and integration of world knowledge. Notably, GPT-4V outperforms LLaVA in task comprehension and better aligns textual intention with visual information.

4.4 Generalization Analysis

Robot Manipulation. To evaluate TEP's generalizability in real robot environments, we extend our method to robot manipulation. As shown in Fig. 5, TEP provides bounding boxes of target objects to a Fetch robot for grasping, following the manipulation framework in (Xiao et al., 2024). Experiments on 30 samples with varied object arrangements and perspectives achieve a success rate of 63.33%, slightly lower than that on the InterRef dataset. Success rates from multiple perspectives hover around 60%, indicating TEP's robust generalization and its effectiveness for HRI.

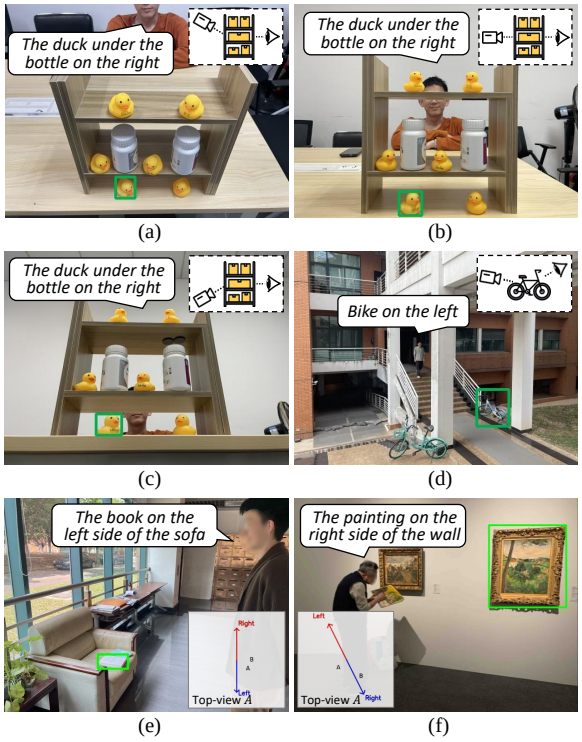


Figure 6: TEP's results across vertical perspectives: (a) high angle, (b) eye level, (c) low angle, and scenarios: (d) outdoor, (e) multiple surfaces, (f) vertical surface.

Vertical Reasoning Capability. Experiments show that TEP extends beyond 2D representations to support effective 3D spatial reasoning.

Vertically Overlapping Objects: In the InterRef dataset, 37.59% of test samples contain objects that overlap along the vertical axis. On these samples, TEP achieves 64.23% accuracy, nearly double the 37.6% accuracy of GPT-4V w/ orientation.

Vertical Spatial Expressions: For expressions that involve vertical relations, TEP achieves 65.22% accuracy versus 26.09% for the baseline. These expressions often require both horizontal and vertical reasoning (*e.g.,* "the book under the box on the left"), which TEP handles effectively by decoupling spatial reasoning into distinct functions.

Vertical Perspective Differences: TEP generalizes across differences in vertical perspectives. As shown in Fig. 6(a-c), within a multi-level shelf where human and robot perspectives differ in vertical angle, TEP effectively captures object information from varying perspectives.

Diverse Surfaces. TEP demonstrates generalization beyond standard tabletop scenarios.

Non-Square Surfaces: 13.44% of samples in the InterRef dataset involves non-square surfaces(*e.g.,* round, oval). TEP achieves 74.45% accuracy on these samples, outperforming the baseline (22.63%). Additionally, in outdoor scenes with stairs and pillars (Fig. 6(d)), adapting the reference plane in the abstract top-view representation from a table to the ground enables robust grounding.

Multiple Surfaces: TEP effectively handles multi-surface scenes, such as books placed on both a sofa and a table (Fig. 6(e)) or paintings on a wall and held in hand (Fig. 6(f)). We introduce a surface identification module that uses VLMs to detect objects on specific surfaces (*e.g.,* sofa, wall). This surface information then informs spatial reasoning and dynamically adapts the top-view representation to relevant surfaces, enabling accurate reasoning.

Vertical Surfaces: Fig. 6(f) illustrates TEP's ability to identify objects on vertical planes. By redefining the top view to align with any object-bearing surface, TEP effectively handles non-horizontal setups, including wall-mounted targets.

5 Error Analysis

We analyze 150 randomly selected predictions and find that 32.00% contain errors, distributed across modules: grounding decomposer (1.33%), object retrieval (3.33%), vertical reasoning (2.00%), and horizontal reasoning (25.33%). Within horizontal reasoning, errors stem from human orientation prediction (4.67%), object position prediction (8.67%), and spatial reasoning using top-view representations (12.00%). Many of these errors stemmed from hallucinations in VLMs, including refusal to respond due to real-person presence (3.33%), misinterpretation of visual details in top views (6.00%), and misinterpretation of prompt cues (5.33%).

6 Conclusion

This paper introduces the HVG task and the InterRef dataset to evaluate perspective-taking in visual grounding. It also presents TEP, a method that enables robots to infer visual information from others' perspectives based on a single self-perspective. The core of TEP is an abstract top-view representation that unifies object positions and human perspectives, allowing robots to understand spatial relationships from different perspectives. It addresses the insensitivity to information differences due to perspective changes. Experiments show that TEP effectively identifies objects from human-centric instructions, generalizes across diverse perspectives, and is applicable to robotic and dynamic scenarios.

Limitations

This paper introduces an HVG task that focuses on human-centric instructions. However, perspective-conditioned instructions vary, such as robot-centric instructions in conventional visual grounding tasks and object-centric instructions. Future work could move forward to disambiguate instructions from diverse roles, perspectives, and multiple speakers.

Besides, this paper mainly focuses on spatial changes caused by perspective shifts, as cognition research (Pearson et al., 2013) predominantly uses spatial information to assess visual perspective-taking abilities. While spatial attributes show significant variation across perspectives, other visual features may also change. For example, a cup with a handle might appear handle-less from a different viewpoint. Future work could explore the impact of perspective on these additional attributes.

Moreover, the proposed method involves LLMs, which currently suffer from slow inference efficiency and may not support real-time robot operations. This limitation could be mitigated by exploring more efficient techniques for LLMs.

Furthermore, while the TEP method achieves approximately 68.29% accuracy on samples involving partially occluded target objects from the robot's perspective, it remains limited in handling fully occluded objects, a challenge even for human cognition. Future work may explore multi-round conversations to address this limitation.

Ethical Considerations

This paper involves human participation through data collection and robot experiments. All individuals in the captured robot scene images and annotation workers have provided informed consent. The COCO images used in the proposed dataset are sourced from a publicly available dataset, ensuring compliance with data usage policies.

To promote transparency and reproducibility, we detail our implementation in Appendix A and will release the code and dataset upon paper acceptance. While the proposed method achieves state-of-the-art performance on the InterRef dataset, we acknowledge that it may not always produce correct predictions. Users should exercise caution when applying the model in critical applications.

Acknowledgments

The authors would like to thank Anxing Xiao for his help with the robot experiments and the reviewers for their valuable comments. This research is supported by the Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078), the National Natural Science Foundation of China (62476097), the Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010), the Fundamental Research Funds for the Central Universities, South China University of Technology (x2rjD2240100), and the China Scholarship Council.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 422–440.

Matt Berlin, Jesse Gray, Andrea Lockerd Thomaz, and Cynthia Breazeal. 2006. Perspective taking: An organizing principle for learning in human-robot interaction. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 1444–1450.

Yuqi Bu, Liuwu Li, Jiayuan Xie, Qiong Liu, Yi Cai, Qingbao Huang, and Qing Li. 2023a. Scene-text oriented referring expression comprehension. *IEEE Trans. Multim.*, 25:7208–7221.

Yuqi Bu, Xin Wu, Liuwu Li, Yi Cai, Qiong Liu, and Qingbao Huang. 2023b. Segment-level and category-oriented network for knowledge-based referring expression comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8745–8757.

Kaiqi Chen, Jing Yu Lim, Kingsley Kuan, and Harold Soh. 2023. Latent emission-augmented perspective-taking (LEAPT) for human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 8006–8013.

Fethiye Irmak Dogan, Sarah Gillet, Elizabeth J. Carter, and Iolanda Leite. 2020. The impact of adding perspective-taking to spatial referencing during human-robot interaction. *Robotics Auton. Syst.*, 134:103654.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 346–355.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Yang Zhao, Xinze Guan, and Xin Wang. 2024. Muffin or chihuahua? challenging multimodal large language models with multipanel VQA. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6845–6863.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, et al. 2024. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 19383–19400.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. 2023. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15326–15337.

Kaixiang Huang, Yuxuan Han, Jinze Wu, Fangzhou Qiu, and Qing Tang. 2022a. Language-driven robot manipulation with perspective disambiguation and placement optimization. *IEEE Robotics Autom. Lett.*, 7(2):4188–4195.

Sihong Huang, Jiaxin Wu, Xiaoyong Wei, Yi Cai, Dongmei Jiang, and Yaowei Wang. 2025. Sound bridge: Associating egocentric and exocentric videos via audio cues. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022b. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 1769–1782.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, and Yu Qiao. 2024. Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22072–22086.

Md Mofijul Islam, Alexi Gladstone, and Tariq Iqbal. 2023. PATRON: perspective-aware multitask model for referring expression grounding using embodied multimodal cues. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 971–979.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1760–1770.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9161–9175.

Roberta L Klatzky. 1998. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*, pages 1–17.

Robyn Langdon and Max Coltheart. 2001. Visual perspective-taking and schizotypy: evidence for a simulation-based account of mentalizing in normal adults. *Cognition*, 82(1):1–26.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965.

Shen Li, Rosario Scalise, Henny Admoni, Stephanie Rosenthal, and Siddhartha S. Srinivasa. 2016. Spatial references and perspective in natural language instructions for collaborative manipulation. In *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016, New York, NY, USA, August 26-31, 2016*, pages 44–51.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. 2024. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 6657–6678.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Zheng Shou. 2024. Exocentric-to-egocentric video generation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, volume 37, pages 136149–136172.

Fangchang Ma, Luca Carlone, Ulas Ayaz, and Sertac Karaman. 2016. Sparse sensing for resource-constrained depth reconstruction. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016*, pages 96–103.

Mark G McGee. 1979. Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological bulletin*, 86(5):889.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large language models as general pattern machines. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pages 2498–2518.

Sabrina M. Neuman, Brian Plancher, Bardienus Pieter Duisterhof, Srivatsan Krishnan, Colby R. Banbury, Mark Mazumder, Shvetank Prakash, Jason Jabbour, Aleksandra Faust, Guido C. H. E. de Croon, and Vijay Janapa Reddi. 2022. Tiny robot learning: Challenges and directions for machine learning in resource-constrained robots. In *4th IEEE International Conference on Artificial Intelligence Circuits and Systems, AICAS 2022, Incheon, Republic of Korea, June 13-15, 2022*, pages 296–299.

Amy Pearson, Danielle Ropar, and Antonia F de C. Hamilton. 2013. A review of visual perspective taking in autism spectrum disorder. *Frontiers in Human Neuroscience*, 7:652.

Pradip Pramanick, Chayan Sarkar, Sayan Paul, Rudra Dev Roychoudhury, and Brojeshwar Bhowmick. 2022. Doro: Disambiguation of referred object for embodied agents. *IEEE Robotics Autom. Lett.*, 7(4):10826–10833.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2022. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3):1623–1637.

Daniel Reich and Tanja Schultz. 2024. Uncovering the full potential of visual grounding methods in VQA. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4406–4419.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13960–13980.

Cheng Shi and Sibei Yang. 2022. Spatial and visual perspective-taking via view rotation and relation reasoning for embodied reference understanding. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 201–218.

J. Gregory Trafton, Alan C. Schultz, Magdalena D. Bugajska, and Farilee Mintz. 2005. Perspective-taking with robots: experiments and models. In *IEEE International Workshop on Robot and Human Interactive Communication, RO-MAN 2005, Nashville, TN, USA, 13-15 August, 2005*, pages 580–584.

Barbara Tversky and Bridgette Martin Hard. 2009. Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, 110(1):124–129.

Jiachen Wang, Zikun Deng, Dazhen Deng, Xingbo Wang, Rui Sheng, Yi Cai, and Huamin Qu. 2025. Empowering multimodal analysis with visualization: A survey. *Comput. Sci. Rev.*, 57:100748.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting

elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Melonee Wise, Michael Ferguson, Derek King, Eric Diehr, and David Dymesich. 2016. Fetch and freight: Standard platforms for service robot applications. In *Workshop on autonomous mobile service robots*, pages 1–6.

Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, Anuja Dawane, Bikramjot Hanzra, Zhuo Deng, Bilan Liu, James Z. Wang, and Cheng-Hao Kuo. 2020. MEBOW: monocular estimation of body orientation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3448–3458.

Xin Wu, Yi Cai, and Ho-fung Leung. 2024. Abstract-level deductive reasoning for pre-trained language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 70–76.

Anxing Xiao, Nuwan Janaka, Tianrun Hu, Anshul Gupta, Kaixin Li, Cunjun Yu, and David Hsu. 2024. Robi butler: Remote multimodal interactions with household robot assistant. *arXiv preprint arXiv:2409.20548*.

Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. 2025. Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. In *International Conference on Learning Representations (ICLR)*, pages 1–15.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. 2023. Universal instance perception as object discovery and retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15325–15336.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 69–85.

A Implementation Details

We optimize prompts and model configurations using the validation set, without requiring fine-tuning of model parameters. The temperatures of the LLMs and VLMs are set to 0, and other configurations default. The generated code is executed via a Python interpreter. To generate the abstract top-view representation, GPT-4V is used to reason and verify object positions, and the Python libraries OpenCV and Pillow are used to draw grids, arrows, and text on a white background. The entire method is deployed on a P100 GPU.

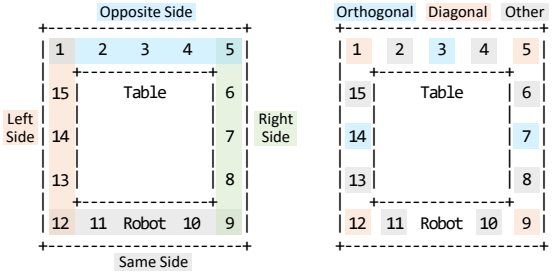


Figure 7: Perspective classifications in the InterRef dataset. (a) Directional perspectives. (b) Angular perspectives.

B Dataset Details

This section supplements Section 4.1 by detailing the data collection, perspective classification, and data statistics of the InterRef dataset.

B.1 Data Collection Details

In the image collection phase, we capture robot scene images from real-world home and school environments, which contain a higher number of similar objects for reference than that of COCO images, establishing an approximate ratio of 2:1 between robot scene and COCO images.

For annotating human-centric instructions (*i.e.,* referring expressions), we employ different approaches based on the image source. In real-world images, we take photos of actual scenes where a person refers to an object and simultaneously record both the referring expression uttered by the person and the corresponding object. For COCO images, we generate referring expressions by imagining the perspective of an individual depicted in the images. After annotation, we assess the quality of each expression and remove samples that involve occlusion. Statistical analysis reveals that most expressions primarily assess perspective-taking through spatial references, while also moderately emphasizing visual attributes such as color.

B.2 Perspective Classification Details

To enhance the classification of perspective types in the InterRef dataset, we label human positions around tables with markers, as depicted by positions 1 to 15 in Fig. 7. Previous studies have classified perspectives into eight categories, *i.e.,* four orthogonal and four diagonal perspectives. However, we find that the observations from positions between an orthogonal position and a diagonal position can also influence the perception of object spatial relationships. Therefore, we supplement eight non-orthogonal and non-diagonal positions. To avoid occlusion, we allocate a specific position for the robot, thereby defining a total of fifteen human positions.

Based on the human position annotations, perspectives could be classified into four directional categories: (1) Opposite side: the human facing the robot across the table (*i.e.,* markers 1 to 5 in Fig. 7(a)); (2) Left side: the human positioned to the robot's left and facing rightward (*i.e.,* markers 1, 12 to 15); (3) Right side: the human positioned to the robot's right and facing leftward (*i.e.,* markers 5 to 9); and (4) Same side: the human positioned on the same side as the robot (*i.e.,* markers 9 to 12). Since minor angular variations can result in perspective changes, being on the same side does not guarantee observing identical object spatial relationships; it only increases the likelihood of a similar perspective. The same principle applies to the other directional categories. Additionally, certain positions may belong to multiple categories. For instance, position 1 falls under both the opposite-side and left-side categories.

Moreover, perspectives can be classified into three angular categories based on whether the person occupies a direct position relative to the robot: (1) Orthogonal: when the person is directly in front, immediately to the left, or immediately to the right of the robot (*i.e.,* markers 3, 7, and 14 in Fig. 7(b)); (2) Diagonal: when the person is positioned at a front-left, front-right, rear-left, or rear-right corner (*i.e.,* markers 1, 5, 9, and 12); and (3) Other: when the position does not fit the aforementioned criteria (*i.e.,* markers 2, 4, 6, 8, 10, 11, 13, and 15).

Method	Test set	Directional perspectives				Angular perspectives		
		Opp.	Left	Right	Same	Orthog.	Diag.	Other
0-shot	31.50	32.15	32.36	26.80	32.19	31.69	26.56	33.81
1-shot	37.25	45.65	37.04	28.21	28.57	39.29	33.33	38.00
3-shot	39.22	43.48	44.44	28.21	42.86	42.86	37.50	38.00
5-shot	39.22	43.48	37.04	30.77	50.00	42.86	37.50	38.00
7-shot	37.25	39.13	40.74	25.64	50.00	42.86	33.33	36.00

Table 6: Results of the baseline method (GPT-4V w/ orientation) in few-shot settings on the InterRef dataset.

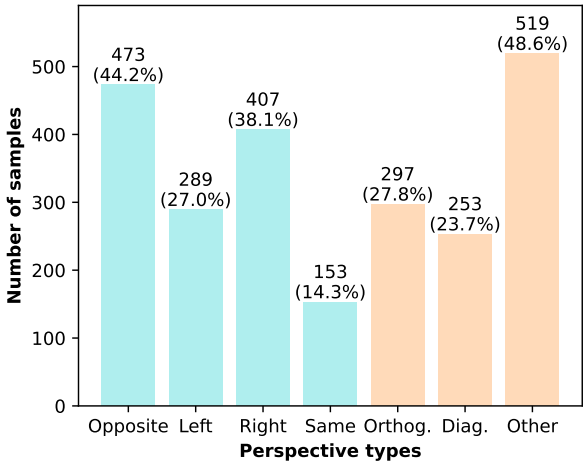


Figure 8: Data distribution of perspective categories.

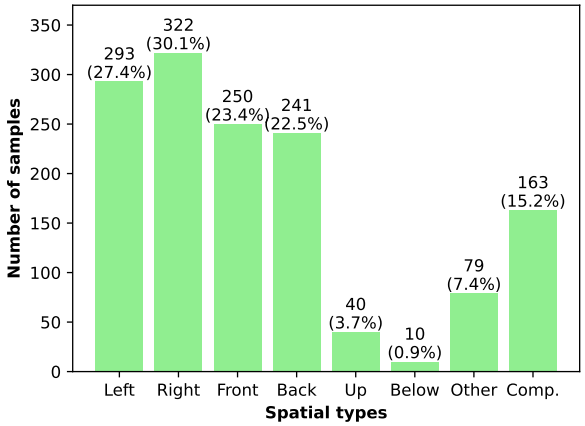


Figure 9: Data distribution of spatial types.

B.3 Data Statistics

Fig. 8 presents the statistical distribution of perspective classifications in the InterRef dataset. For directional categories, the majority of samples fall into the "opposite" category, while the "same" category has the fewest samples. This distribution reflects the perspective imbalance issue discussed in Section 2.2, where face-to-face human-robot interactions are predominant, limiting the visual perspective-taking capabilities of foundation models. Additionally, in angular categories, orthogonal perspectives are more frequent than diagonal perspectives. Other types of angular perspectives collectively account for approximately 50% of the dataset, which corresponds to the fact that positions in this category constitute half of the total.

Fig. 9 illustrates the data distribution of spatial types mentioned in human-centric instructions. The "left" category includes instructions that contain leftward-related terms such as "leftmost" and "left-handed," while the "back" category encompasses terms like "behind" and "rear." Similar classifications apply to other directional categories. Additionally, the "other" category represents spatial references beyond standard horizontal and vertical directions, including terms like "next," "between," and "center." Instructions that involve multiple spatial relationships are categorized under "Composition (Comp.)." The statistical analysis reveals that instructions requiring horizontal reasoning (*i.e.,* left, right, front, and back) are the most prevalent. Meanwhile, instructions categorized as "other," which involve more complex spatial reasoning, constitute approximately 7.4% of the dataset. Furthermore, instructions in the "Composition" category, which necessitate multi-step spatial reasoning, account for 15.2%.

C Few-Shot Settings

To ensure a fair comparison, the baseline VLM (GPT-4V w/ orientation) receives the same visual inputs as our method: images, object bounding boxes, depth values, and human orientation angles. Given the VLMs' ability to infer perspective-taking in a zero-shot manner, we adopt a zero-shot evaluation setting.

We also include a few-shot baseline (Table 6) to further assess performance. Prompts are enhanced using in-context learning, with 5-shot and 7-shot settings incorporating examples across four directional and three angular perspectives.

Experiments show that TEP in the zero-shot setting outperformed the baseline in the few-shot setting. While in-context learning improves performance over the zero-shot baseline, gains diminish after 3-shot, with performance stabilizing or declining at 5-shot. This decline may be due to: (1) the complexity of example scenarios, where reasoning varies even within the same perspective, and (2) the limitations of textual CoT examples in capturing cross-perspective spatial reasoning. In contrast, the abstract top-view representation in TEP provides a more intuitive encoding of spatial information across perspectives.

D More Qualitative Results

D.1 Comparison with Baselines

Fig. 10 presents the visualization results comparing the TEP method with the baseline model, GPT-4V, which incorporates human orientation. The responses from GPT-4V indicate that, to some extent, the baseline model can infer the human perspective based on the provided orientation. However, it struggles with spatial reasoning, often misinterpreting spatial relationships within the human perspective, highlighting its insensitivity to information differences across perspectives.

For example, in Fig. 10(a), the baseline model incorrectly equates "closest to the robot" with "closest to the person," failing to distinguish between these perspectives. In contrast, the TEP method, leveraging a generated top-view representation, accurately identifies the third cake as being in front when observed from the human perspective. Similarly, in Fig. 10(b), the baseline model incorrectly assumes that a pizza farther from the robot's perspective is also positioned at the back from the human perspective. In contrast, the TEP method systematically identifies the target object by applying vertical and horizontal reasoning based on the abstract top-view representation. These visualization results demonstrate the effectiveness of TEP in visual perspective-taking, showcasing its ability to perform human-centric visual grounding more accurately than the baseline model.

D.2 Dynamic Scenario

We explore the applicability of the abstract top-view representation in dynamic scenarios. As shown in Fig. 11, TEP adapts to the movements of humans, objects, and robots by updating the top-view information, *e.g.,* by reorienting a human direction, repositioning an object, or rotating the view to align with the current direction. These results show that TEP is applicable in dynamic scenarios due to the flexibility of top-view representations supporting partial updates. Future work could involve a monitoring mechanism to trigger these updates.

D.3 Distance from Tables

According to geometric optics and linear projection, distance affects the perceived density of objects but not their spatial arrangement, provided the perspective remains unchanged. As shown in Fig. 12, we conduct experiments at distances ranging from 0 to 5 meters. Results show that distance has a minimal effect on the orientation prediction network, with orientation angle variations remaining within 5 degrees. This corresponds to a consistent human position in the top-view representation, resulting in stable and accurate predictions.

E Robot Experiment Details

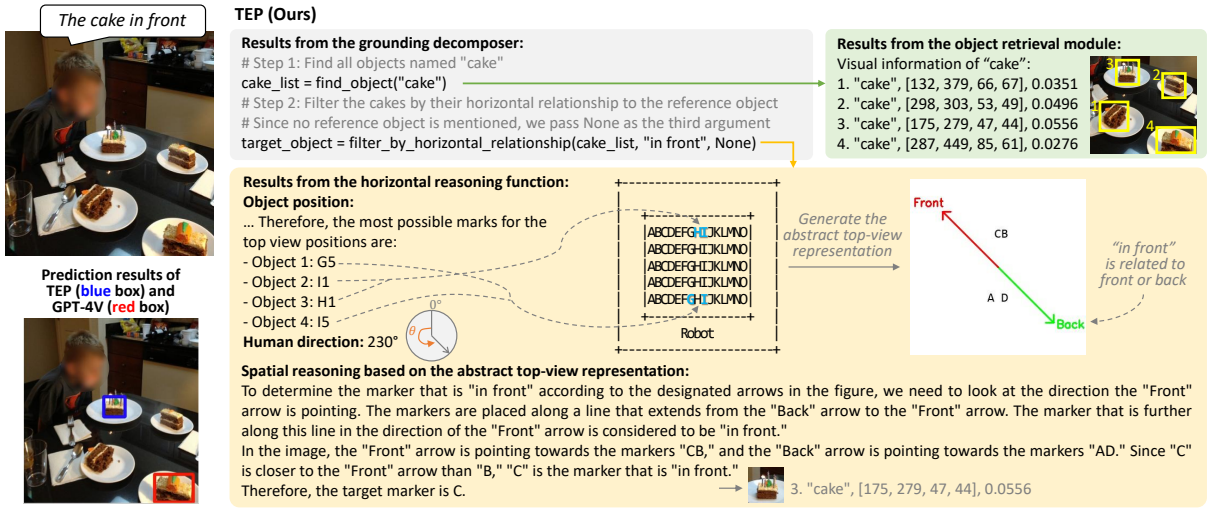
In our robot experiments, we utilize a Fetch mobile manipulator (Wise et al., 2016) equipped with a camera, arm, and gripper. The experiments involve physical interactions with seven distinct object types, such as bottles, tape rolls, and chip bags, representing a range of shapes, sizes, weights, and materials. To ensure diverse testing conditions, we vary the object arrangements as well as the relative positions between humans and the robot, covering all perspective categories. Human-centric instructions are provided by four volunteers. The experimental results indicate a success rate of 63.33%, which is slightly lower than that achieved on the InterRef dataset. While TEP is not restricted to specific object types, the robot's grasping performance is constrained by the mechanical limitations of its arm. Please refer to the supplemental materials for a demonstration video of the robot experiments.

F Prompt Constuction

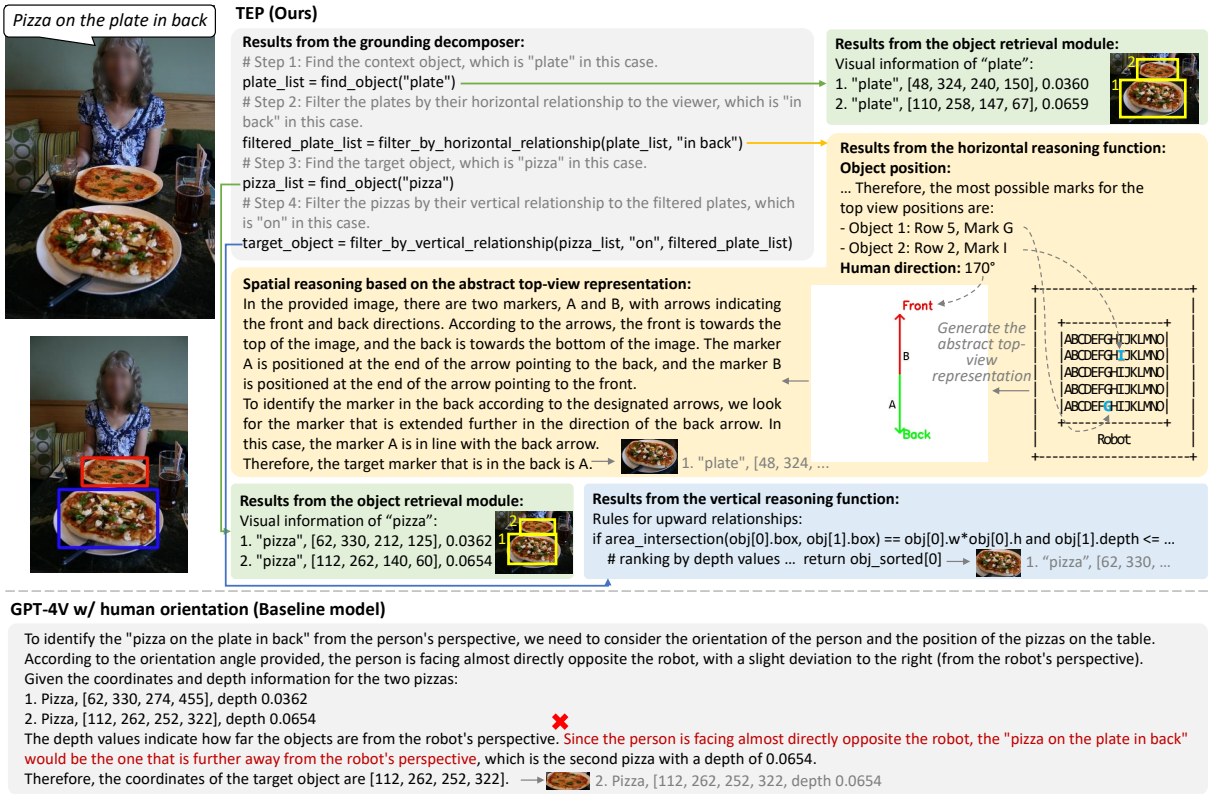
TEP employs prompts to elicit LLMs and VLMs to decompose the grounding task, construct an abstract top-view representation, and reason about spatial relationships based on this representation.

F.1 Prompt for Grounding Decomposer

Fig. 13 shows the prompt for the grounding decomposer. Given a referring expression, an LLM is prompted to decompose the task into a series of



(a)



(b)

Figure 10: Visualization of the TEP method and comparison with the baseline model, GPT-4V w/ human orientation.

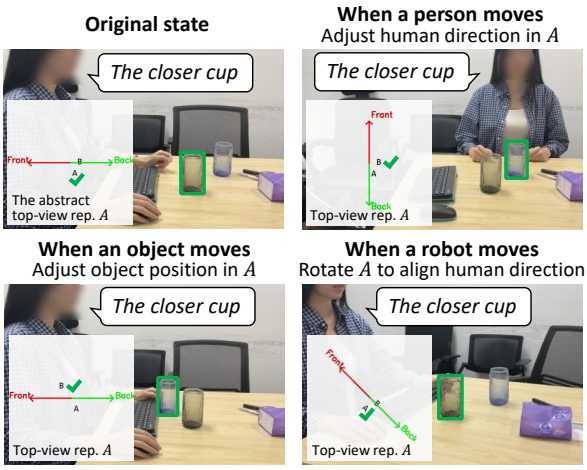


Figure 11: Visualisation of top view updates and final results (depicted in green boxes) in dynamic scenarios.

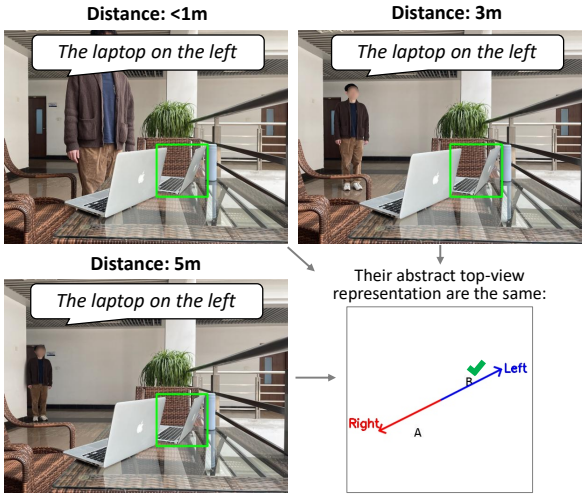


Figure 12: TEP's results (green boxes) at varying distances from tables.

grounding steps employing a context-first strategy and generate programs associated with these steps.

F.2 Prompt for Top-View Representation

Fig. 14 shows the prompt for reasoning about object positions in the top-view representation. We provide visual information about objects from the robot's perspective, including bounding boxes and depth values. The VLMs assign each object a position within the top view wherein the robot's position serves as a landmark.

When object positions are found to be incorrect, the proposed method elicits the VLMs to revise them in a second round. The prompt is shown in Fig. 15, which includes an error description, such as "since the depth value of object A is greater than that of object B, object A might be above object B in the top view."

When the human direction is found to be incorrect, we elicit the VLMs to reason about the human position in the top view based on the visual information from the robot's perspective. The prompt is illustrated in Fig. 16.

The prompt for horizontal reasoning based on the abstract top-view representation is shown in Fig. 17. The proposed method prompts the VLMs to identify objects that satisfy a constraint in a given relational phrase, specifically in relation to designated directions.

F.3 Prompt for Baseline Methods

As shown in Fig. 18, the prompt for baseline models provides a detailed task background and visual information about objects from the robot's perspective. In addition, for a set of GPT-4V experiments, the angle of human orientation is additionally provided. This information is then utilized to prompt models to identify target objects as referred to in a referring expression from a human perspective.

Call the three optional external functions provided below to decompose a visual grounding task with an input sentence into several steps, avoid using any other functions or properties, and avoid generating code to process data.

```
def find_object(name: str, attribute: str=None) -> List[Object]:
    # Find all objects with the given name and visual attribute (optional, e.g., color, shape, texture). This function utilizes an object detector and does not support spatial reasoning. This function outputs a list of objects.
    pass

def filter_by_vertical_relationship(object_list: List[Object], subsentence: str, reference_object_list: List[Object]=None) -> List[Object]:
    # Filter objects by their vertical relationship (e.g., above, below) to the reference objects (optional). This function outputs a list of objects that satisfy a subsentence.
    # Example: box on top -> filter_by_vertical_relationship(box_list, "on top", None)
    # Example: box on the computer -> filter_by_vertical_relationship(box_list, "on", computer_list)
    pass

def filter_by_horizontal_relationship(object_list: List[Object], subsentence: str, reference_object_list: List[Object]=None) -> List[Object]:
    # Filter objects by their horizontal relationship (e.g., left, right, front, back, nearest, far, right rear) to the reference objects (optional). This function outputs a list of objects that satisfy a subsentence.
    # Example of subsentence: second from left, between, ...
    # Example: box on the left -> filter_by_horizontal_relationship(box_list, "on the left", None)
    # Example: the first box to the left of computer -> filter_by_horizontal_relationship(box_list, "first to the left of", computer_list)
    pass
```

In this case, the input sentence is "<REFERRING_EXPRESSION>". If the sentence mentions multiple objects, you should use a top-down and context-first approach to decompose it into multiple steps. Begin with context objects (and any related spatial reasoning) and progressively narrow down to the target object (and any related spatial reasoning). Include only the necessary steps in the code, and assign the final result to a variable named `target_object`.

Figure 13: Prompt for decomposing the task into a series of grounding steps in the grounding decomposer. The highlighted section is adjusted according to the given information.

A simulated human-robot interaction scenario involving a robot and a person facing and surrounding a table. The input image is captured by the robot. We construct an estimated top view using ASCII art representation, where the central rectangle indicates the table, and the mark 'Robot' indicates the position of the robot around the table, respectively. There are 5 rows of marks 'A' to 'O' in the central rectangle that indicate the possible positions of objects on the table.

```
  +----------------------+
  |                      |
  |   +---------------+  |
  |   |ABCDEFGHIJKLMNO|  |
  |   |ABCDEFGHIJKLMNO|  |
  |   |ABCDEFGHIJKLMNO|  |
  |   |ABCDEFGHIJKLMNO|  |
  |   |ABCDEFGHIJKLMNO|  |
  |   +---------------+  |
  |        Robot         |
  +----------------------+
```

We focus on specific objects on the table captured by the robot's view at x-coordinates of center points, along with their depth values (where a smaller depth value indicates objects closer to the camera):
<OBJECT_INFORMATION>

You should determine the position of these objects in the estimated top view. First, analyze spatial relationships among the objects. Second, analyze relative regions of the objects on the table, where the table x-coordinates are <TABLE_X>. In the horizontal direction, objects on the left half of the table (x-coordinates <TABLE_X_MIN>) correspond to marks 'A' to 'G' in the top view, on the right half of the table (x-coordinates <TABLE_X_MAX>) correspond to marks 'I' to 'O', and in the horizontal middle of the table correspond to mark 'H'. In the vertical direction, objects on the upper half of the table's surface correspond to rows 1 to 2, on the bottom half correspond to rows 4 to 5, and in the vertical middle correspond to row 3. You can use x-coordinate values to infer horizontal relationships. Use depth values and analyze the table surface in the input image to infer vertical relationships.

Finally, integrate these analyses to identify the exact row number and column mark of each object's top view position. Ensure that each position accommodates a maximum of one object. Output with the statement 'Therefore, the most possible marks for the top view positions are: - Object 1: ... - Object ...'

Figure 14: Prompt for reasoning about the positions of objects in a top view and assigning rows and lines to each object's position. The highlighted section is adjusted according to the given information.

Your answer may be incorrect, because it does not match the following object relationships:
<ERROR_DESCRIPTION>

Please reanalyze the position of the objects after considering the above factors and output the final answer.

Figure 15: Prompt for revising the collected object positions in the verification process. The highlighted section is adjusted according to the given information.

A simulated human-robot interaction scene involving a robot and a person facing and surrounding a table. The input image is captured by the robot. The coordinates [x_min, y_min, x_max, y_max] and depth values (smaller depth value corresponds to objects closer to the robot) of the person and the table are as follows:
<OBJECT_INFORMATION>
Based on this scene, we construct an estimated top view using ASCII art representation, where the central rectangle indicates the table, and the robot, i.e., the scene shooter, positions at the bottom center. Number marks 1 to 15 in the top view represent the possible positions of the person.

```
  +----------------------+
  | 1    2    3    4    5 |
  |   +--------------+   |
  | 15|     Table    | 6 |
  |   |              |   |
  | 14|              | 7 |
  |   |              |   |
  | 13|              | 8 |
  |   +--------------+   |
  | 12  11  Robot  10   9 |
  +----------------------+
```

To determine the person's position, you should firstly understand the input image and use depth values to analyze whether the person is on the opposite side of the table, on the left/right side of the table, or on the same side of the table as the robot.

Secondly, you should understand the input image to further analyze the person's position. You can use x-coordinate values to infer horizontal relationships and determine the position of the person's center point x_center=<PERSON_X_MID> in the horizontal direction.
1. If the person is on the opposite side of the table:
- If the front of the person is fully visible and the person is directly across the table, it corresponds to mark 3, while slightly to the left corresponds to mark 2 and slightly to the right corresponds to mark 4.
- If the person is diagonally facing the table, in this case if the person is in the left rear corner it corresponds to mark 1, while if the person is in the right rear corner it corresponds to mark 5.
2. If the person is on the left/right side of the table:
- If the side of the person is observable and the person is directly on the right of the table, it corresponds to mark 7. In this case, if the front of the person is partially visible, it corresponds to mark 6, while if the back of the person is partially visible, it corresponds to mark 8.
- If the side of the person is observable and the person is directly on the left of the table, it corresponds to mark 14. In this case, if the front of the person is partially visible, it corresponds to mark 15, while if the back of the person is partially visible, it corresponds to mark 13.
3. If the person is on the same side of the table as the robot:
If the person is in the right corner, in this case if the back of the person is partially visible it corresponds to mark 9, while if the back is fully visible it corresponds to mark 10.
If the person is in the left corner, in this case if the back of the person is partially visible it corresponds to mark 12, while if the back is fully visible it corresponds to mark 11.

Finally, integrate the preceding analysis results and present a conclusion, stating 'Therefore, the most possible number mark for the person's position in the top view would be ...'

Figure 16: Prompt for reasoning about human positions in a top view within the verification process. The highlighted section is adjusted according to the given information.

According to the given information and your knowledge, answer the question.

Your task is to identify the marker(s) <RELATIONAL_PHRASE> in the input figure using the designated <ASSOCIATED_DIRECTIONS> arrows in the figure. First, interpret "<RELATIONAL_PHRASE>" relative to the intended orientation. For example, "rear" and "behind" align with the back direction, while "close" and "near" align with the front direction. Then, disregard conventional <ASSOCIATED_DIRECTIONS> directions, focusing solely on the designated arrows within the figure. The letter that extends further in the direction of the arrow conforms more to a specific orientation. Output and assign the target marker(s) to the variable target_marker = .

Let's think step by step.

Figure 17: Prompt for horizontal reasoning based on the abstract top-view representation. The highlighted section is adjusted according to the given information.

A simulated human-robot interaction scenario involving a robot and a person facing and surrounding a table. The input image is captured by the robot, containing the following objects with coordinates [x_min, y_min, x_max, y_max] and depth (smaller value means closer to the robot):
<OBJECT_INFORMATION>
The orientation angle of the person is <ORIENTATION_ANGLE>, which is obtained from a human orientation estimation model, where 0, 90, 180, and 270 degrees represent the person on the same, right, opposite, and left sides of the table, respectively.

When the person refers to an object "<REFERRING_EXPRESSION>" from the person's perspective, you should identify the target object and output "Therefore, the coordinates of the target object are ...". Let's think step by step.

Figure 18: Prompt for baseline models. The highlighted section is adjusted according to the given information, and the blue text is human orientation information additionally provided to a set of GPT-4V experiments.